Hydropathy profile alignment

Lolkema, JS; Slotboom, DJ

Published in:
FEMS Microbiology Reviews

DOI:
10.1111/j.1574-6976.1998.tb00372.x

Publication date:
1998

ELSEVIER

FEMS Microbiology Reviews 22 (1998) 305–322

FEMS MICROBIOLOGY REVIEWS

Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins

Juke S. Lolkema *, Dirk-Jan Slotboom

Department of Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands

Received 25 May 1998; received in revised form 17 August 1998; accepted 22 August 1998

Abstract

Hydropathy profile alignment is introduced as a tool in functional genomics. The architecture of membrane proteins is reflected in the hydropathy profile of the amino acid sequence. Both secondary and tertiary structural elements determine the profile which provides enough sensitivity to detect evolutionary links between membrane proteins that are based on structural rather than sequence similarities. Since structure is better conserved than amino acid sequence, the hydropathy profile can detect more distant evolutionary relationships than can be detected by the primary structure. The technique is demonstrated by two approaches in the analysis of a subset of membrane proteins coded on the *Escherichia coli* and *Bacillus subtilis* genomes. The subset includes secondary transporters of the 12 helix type. In the first approach, the hydropathy profiles of proteins for which no function is known are aligned with the profiles of all other proteins in the subset to search for structural paralogues with known function. In the second approach, family hydropathy profiles of 8 defined families of secondary transporters that fall into 4 different structural classes (SC-ST1–4) are used to screen the membrane protein set for members of the structural classes. The analysis reveals that over 100 membrane proteins on each genome fall in only two structural classes. The largest structural class, SC-ST1, correlates largely with the Major Facilitator Superfamily defined before, but the number of families within the class has increased up to 57. The second large structural class, SC-ST2 contains secondary transporters for amino acids and amines and consists of 12 families.

Keywords: Hydropathy profile alignment; Membrane protein structure; Structural classification; Secondary transporter; Major Facilitator Superfamily; Functional genomics

Contents

* Corresponding author. Tel.: +31 (50) 3632155; Fax: +31 (50) 3632154; E-mail: j.s.lolkema@biol.rug.nl

0168-6445/98/$19.00
PII: S0168-6445(98)00018-7

1. Introduction

The genomes that have been sequenced completely today represent a massive amount of data and it may be expected that with the ongoing genome sequencing projects, this amount will grow rapidly. Analysis of the data will be a major task in the near future. The greater part of the ORFs on the genomes available today code for proteins of unknown functions. It is impossible to look for these functions by experimental techniques without having a clue, and, therefore, the first step in functional genomics is the search for homologues in the available databases. Alignments of the amino acid sequences that result in significant identity reveal a common evolutionary origin, and, possibly, a similar function.

In the course of evolution, the structure of proteins is better conserved than the amino acid sequence of the polypeptide chain. Apparently, 3D-structure is quite tolerant to changes in primary structure. Consequently, when two proteins have diverged in evolution, the evolutionary relationship will be longer detectable from a comparison of the three dimensional fold of the proteins than from the two amino acid sequences. Obviously, in the practical sense this is not very useful because it is not easy to obtain the structures of the proteins. The solution would be to predict the structure of proteins by computational techniques based on the amino acid sequences, but this is not yet possible. In this context, a special class of proteins is formed by integral membrane proteins. Because of their simple architecture, the hydropathy profile of the amino acid sequence [1] provides a fingerprint of their structure. Similar to structure, the hydropathy profile of a membrane protein is better conserved than the amino acid sequence from which it is calculated and the hydropathy profiles have been used to demonstrate the evolutionary relationship between different families of membrane proteins [2].

It is expected that hydropathy profile alignment techniques will identify more distant evolutionary relationships between membrane proteins than amino acid sequence alignments. They may provide an additional tool for the identification of the function of the proteins coded by the many ORFs on a genome. In this paper, the relation between the structure of membrane proteins and the hydropathy profiles of the amino acid sequence will be reviewed, the techniques to find the optimal alignment will be discussed briefly and the prospects of using this tool in functional genomics projects will be given.

2. Membrane protein structure

2.1. A bundle of α-helices

Even though the three-dimensional structure of only a handful of membrane proteins has been determined, it is believed that membrane proteins of the plasma membrane of bacterial and eukaryotic cells, as well as of organelles, share a similar architecture that is characterized by a single secondary

structure element: the α-helix. The proteins consist of a bundle of α-helices that is oriented perpendicular to the plane of the membrane. The number of helices may differ from 2 up to as many as 17. They are connected by loops that vary considerably in length from a few residues to loops long enough to fold in domain-like structures in the periphery of the membrane. The loops contact the water phase or may provide attachment sites for extra-membrane subunits of multi-component assemblies. Similarly, in the membrane, the transmembrane helices contact the lipid phase or interact with other integral membrane subunits. Therefore, in spite of their similar architecture, membrane proteins form many different structures designed to perform as many different functions.

2.2. Conservation of membrane protein structure

The structures of three pairs of membrane proteins are available to show the conservation of the 3D-structure within a family of homologous proteins. The photosynthetic reaction centers of the purple bacteria *Rhodopseudomonas viridus* and *Rhodobacter sphaeroides* were the first membrane proteins to be crystallized and the structures were determined at atomic resolution [3–5]. The reaction center that functions as a light driven electron pump is a complex assembly of multiple proteinaceous subunits and a multitude of pigment molecules. The core of the membrane embedded part is formed by the homologous subunits L and M that fold similarly as five helix bundles. Two helices of each subunit interact intimately to form a four helix bundle motif resulting in a pseudo 2-fold rotational symmetry axis perpendicular to the plane of the membrane. Between the helices the pigment molecules are bound. The L and M subunits of the reaction centers of *R. viridus* and *R. sphaeroides* are genetically homologous proteins (Table 1) and their three dimensional structures are almost superimposable even though their amino acid sequence identity may be as low as 27%.

Cytochrome *c* oxidases catalyze electron transfer from reduced cytochrome *c* to oxygen while conserving the free energy by pumping protons across the membrane. The enzyme complexes are widely distributed throughout nature and are found in bacteria, archaea and lower and higher eukarya suggesting large evolutionary distances. The latter is apparent from the much more complicated subunit structure of the eukaryotic complexes when compared to the bacterial ones. The cytochrome *c* oxidases of the bacterium *Paracoccus denitrificans* and bovine heart mitochondria have been crystallized and the structures have been determined at a resolution of 2.8 Å [6,7]. The part of the structures formed by subunits I, II and III that is essential to catalysis and shared by both enzymes, shows a similar folding. The central catalytic subunit I folds as a 12 helix bundle with a

Table 1
Divergence of amino acid sequences and conservation of structure

	pufLrhs	pufLrv	pufMrhs	pufMrv
pufLrhs	–	59	30	28
pufLrv		–	33	27
pufMrhs			–	49
pufMrv				–

Indicated are the pairwise amino acid sequence identities of the pufL and pufM subunits of *Rhodobacter sphaeroides* (rhs) and *Rhodopseudomonas viridis* (rv) reaction centers. The two subunits form the core of the protein complexes. The crystal structures of the complexes reveal an almost identical three dimensional folding of the L and M subunits in 5 transmembrane α-helices. Sequences were aligned using Clustal W [11] with the standard setting of the parameters. Sequence identity is defined as the number of identical residues in the alignment divided by the length of the shortest sequence.

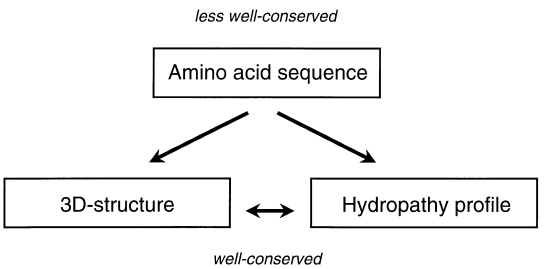


Fig. 1. The amino acid sequence, 3D-structure and hydropathy profile on an evolutionary scale. The hydropathy profile of a membrane protein is intermediate between the amino acid sequence and the three dimensional structure of the protein; the profile is calculated from the primary structure and is a fingerprint of the 3D-structure. On an evolutionary time scale hydropathy profiles evolve at a similar rate as 3D structures which is slower than the divergence of primary structures. Consequently, the hydropathy profile contains early evolutionary links that may be lost in the amino acid sequence.

pseudo 3-fold rotational symmetry axis perpendicular to the plane of the membrane and incorporates the two haem groups and the copper B center. The sequence identity between the *Paracoccus* and mitochondrial subunits I is 53%. Subunits III that fold as a 7 helix bundle share 50% sequence identity.

In the light driven proton pump bacteriorhodopsin of *Halobacterium salinarium* the retinal cofactor is positioned halfway the membrane in a bundle of seven transmembrane helices [8]. Bacterio-opsin, the precursor of bacteriorhodopsin, shares 30% sequence identity with halo-opsin, the precursor of halorhodopsin, a light driven chloride pump. The structure of both retinal proteins has been determined from the analysis of two dimensional crystals formed in the plane of the membrane which has resulted in less well resolved structures with 3.5 and 7 Å resolution for bacteriorhodopsin and halorhodopsin, respectively. At this level of resolution the fold of the two proteins is similar [8,9].

These three examples do not only show the conservation of structures within homologous families of membrane proteins but also the insensitivity of the folding of the polypeptide chain to changes in the amino acid side chains. This stresses that during evolution amino acid sequences diverge much faster than the structures for which they code (Fig. 1).

2.3. A fingerprint of membrane protein structure

The repeated helix-loop-helix motif in the structure of membrane proteins is reflected in the hydropathy profile of the amino acid sequence. The transmembrane α-helices or transmembrane segments (TMS) span the hydrophobic core of the phospholipid bilayer and, for energetic reasons, have an overall high hydrophobicity themselves. The connecting loops, that are in contact with the water phase or the hydrophilic surface of other subunits are likely to be

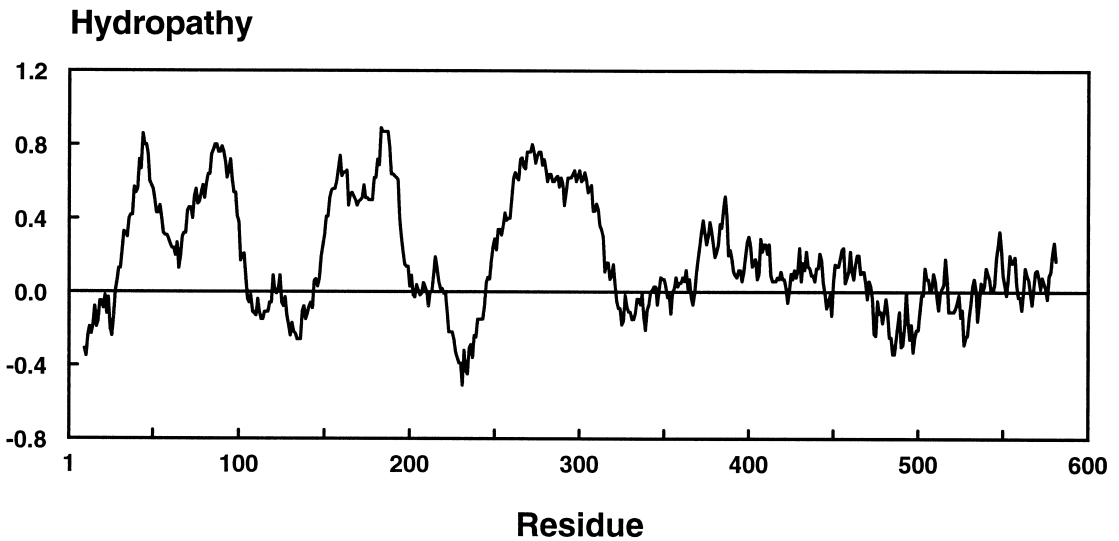


Fig. 2. The hydropathy profile is a fingerprint of membrane protein structure. Shown is the hydropathy profile of LmrA, a multi drug transporter of *Lactococcus lactis* and bacterial homologue of P-glycoprotein [35]. The protein belongs to the ATP driven ABC transporters and has a typical two domain structure. The N-terminal half of the protein up to about residue 325 is an integral membrane domain that is responsible for the transport activity of the protein. The C-terminal half of the protein, the ABC domain, is protruding into the cytoplasm and has the characteristics of a soluble protein. The ABC domain is the energy coupling domain that allows the transport of the substrates against their concentration gradient at the expense of ATP hydrolysis. The typical structure of the membrane bound domain is reflected in the hydropathy profile of the N-terminal part, that is characterized by a number of hydrophobic peaks. The profile of the ABC domain shows that a soluble protein has a much less 'structured' hydropathy profile. The profile of the membrane bound domain reveals three pairs of transmembrane α-helices (residues 30–100, 150–200 and 250–320) separated by two relatively hydrophilic regions. The resulting structure would be a 6 helix bundle with connecting loops that are longer at the cytoplasmic side of the membrane than at the extracellular side of the membrane. The hydropathy profile was computed using the hydrophobicity scale of Eisenberg [36] and a window of 19 residues.

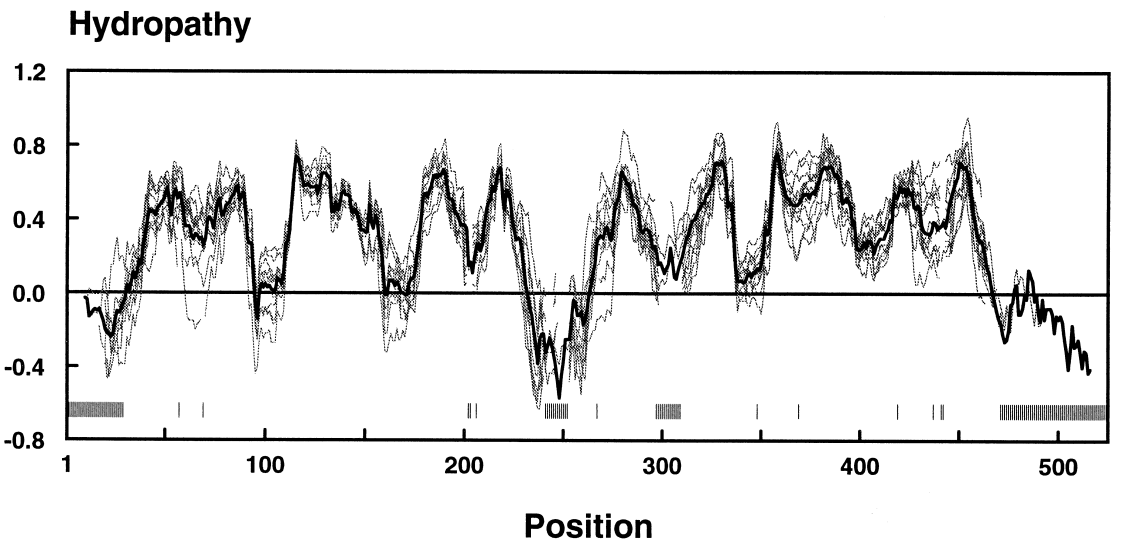


Fig. 3. Hydropathy profile of a family of homologous membrane proteins. The averaged hydropathy profile of a family of 12 citrate, α-ketoglutarate and proline transporters (CitKgl) and the individual profiles of the members are indicated in bold and as thin lines, respectively. Regions of high and low hydrophobicity coincide in all the individual profiles, which results in a family profile that shows the same pattern as each of the members. The family profile averages out the noise in the individual profiles and is the best fingerprint of the folding of the polypeptide chain that is common to all members. The distribution of the individual profiles around the family profiles provides a measure for the divergence within a homologous family which is expressed in a parameter termed the structure divergence score (SDS). The SDS is defined as the average distance between the family and individual profiles at each position and equals 0.117 hydrophobicity units for this family. The family members were given in [2] and aligned using the Clustal W program [11]. Pairwise sequence identities in the aligned set ranged between 22 and 68% with a median of 28%. Vertical bars indicate positions in the alignment where gaps occur in any of the sequences. Typically, these gaps occur in the hydrophilic regions, representing the loops in the protein structure.

more hydrophilic. This results in stretches of alternating hydrophobic and hydrophilic residues in the amino acid sequence of the proteins which is reflected in the typical peaks in the hydropathy profile of a membrane protein (Fig. 2). The hydrophobic core of the phospholipid bilayer has a thickness of 30 Å and it takes about 20 hydrophobic residues to span this distance in an α-helical conformation. The length of the loops is not restricted and there length is much more variable. The length of the loops largely determines how well the number of TMSs in the structure of the protein is resolved in the hydropathy profile. With sufficiently large loops, the number of α-helices is immediately evident from the hydropathy profile, i.e. a simple analysis of the amino acid sequence results in a secondary folding model of the polypeptide chain in the membrane, an analysis that is much more cumbersome in the case of globular proteins. Furthermore, there is a statistical bias towards positively charged residues in cytoplasmic loops ('positive inside' rule) which allows the determination of the orientation of the protein in the membrane [10].

As the hydropathy profile of membrane proteins is a reflection of the structure of the protein, it may be expected that the hydropathy profile will be better conserved than the amino acid sequence from which it is calculated. This is evident from multiple sequence alignments of homologous membrane proteins [11] that can be used to compute an averaged hydropathy profile for the family. The pattern of hydrophobic peaks is more or less the same in all members and coincides with the average profile even though pairwise sequence alignments between the members are as low as 20% (for an example see Fig. 3) [2]. The family profile provides a better fingerprint of the structure of the members of the family since it averages out the noise in the individual profiles. We have used the family profiles to show the evolutionary relationship between secondary transporter families of which the members could not be demonstrated to be homologous by

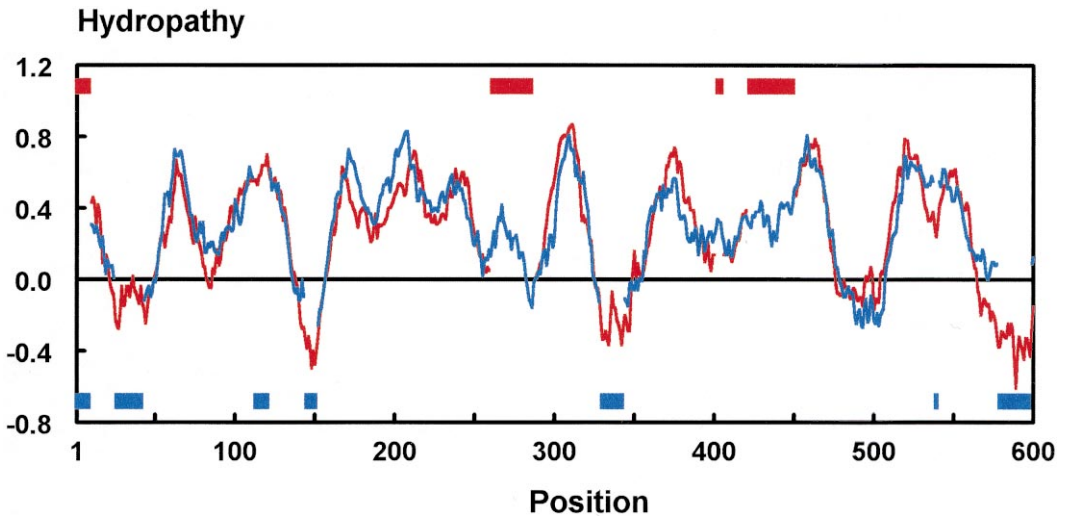


Fig. 4. Structural similarity of the SecY and Sec61α protein families. The SecY and Sec61α integral membrane proteins are the largest subunit of the bacterial, archaeal and eukaryotic preprotein export machinery. They are believed to be 10 helix bundles that together with two smaller subunits form the core of a hydrophilic pore through the membrane (the translocon) through which the nascent polypeptide chains are transported across the membrane (reviewed in [37,38]). The SecY family contains 18 proteins from bacterial origin, while the Sec61α family contains proteins from eukaryotes and archaea. Members of the two families are only distantly related based on amino acid sequence. For instance, the first member of the Sec61α family picked up by a BLAST search using the *E. coli* SecY as the query is the preprotein translocase of *Methanococcus jannaschii*. The two proteins share 30% sequence identity in a total of 204 residues fragmented over 7 stretches ($P = 1.9e{-}7$). The optimal alignment of the family hydropathy profiles of the SecY family (red) and the Sec61α family (blue) reveals a very similar pattern strongly suggesting a similar folding. The gaps indicated at the top and bottom are mostly in the hydrophilic loop regions. The SecY family is characterized by an SDS of 0.138 [2]. The Sec61α family contained 7 members from *Haloarcula marismortui* (accession no. P28542), *Methanococcus vaviellii* (P28541), *Pyrenomonas salina* (P38379), *Rattus norvegicus* (P38378), *Sulfolobus acidocaldarius* (P32915), *Saccharomyces cerevisiae* (P38353). The pair wise sequence identity of the aligned sequences ranged between 22 and 62% and the SDS was 0.105.

amino acid sequence identities [2]. Optimal alignment of the family profiles showed that, in fact, they were very similar. Fig. 4 gives another example of structural similarity of two families of membrane proteins to demonstrate that hydropathy profile alignments can detect more distant evolutionary relationships than amino acid sequence alignments (Fig. 1).

2.4. Structural resolution of hydropathy profiles

The hydropathy profile of the amino acid sequence easily discriminates between membrane proteins and globular proteins and, in many cases, the hydropathy profile of membrane proteins discriminates between proteins containing different numbers of transmembrane helices. The structure of a membrane protein is largely determined by the way the transmembrane helices are folded relatively to one another. In contrast to the secondary structure information, the tertiary structure cannot be deduced from the hydropathy profiles, but it seems that the tertiary folding, at least in part, does define the profile. The family hydropathy profile reflects both the secondary and tertiary structure elements that form the skeleton that is common to the members of the family. This means that different family structures with the same number of transmembrane segments can be distinguished on the basis of their hydropathy profile. Discrimination between two family hydropathy profiles is based on the diversion observed in the individual hydropathy profiles of the members of a homologous family that reflect the same structure [2]. Two families are said to share the same global structure when the difference between their family hydropathy profiles is about the same as the average difference between the individual hydropathy profiles within the two families (see below). Using this criterion, 8 dif-

ferent families of secondary transporters, membrane proteins that typically consist of 12 transmembrane segments, could be classified in 4 different structural classes (Structural Class-Secondary Transporters 1–4, SC-ST1–4). SC-ST1 contained 4 families: the monosaccharide symporters and uniporters (Sugar in Fig. 5), di- and tricarboxylate symporters (CitKgl), drug antiporters (Tetracyc) and disaccharide symporters (GPH). In its original definition, these families were all in the Major Facilitator Superfamily [12,13]. The second class, SC-ST2, contained 2 families: an abundant family of amino acid symporters (AmAc) and the Na^+-dependent neurotransmitter symporters (SNF). SC-ST3 and SC-ST4 each contained a single family, the glutamate symporters (Glus) and the gluconate symporters (Gluconat) [2]. In this contribution, these families will be used to make a structural classification of the secondary transporters coded on the *E. coli* and *B. subtilis* genomes. A schematic representation of the structural classes and homologous families with a few representative transporter proteins of the two bacteria is presented in Fig. 5.

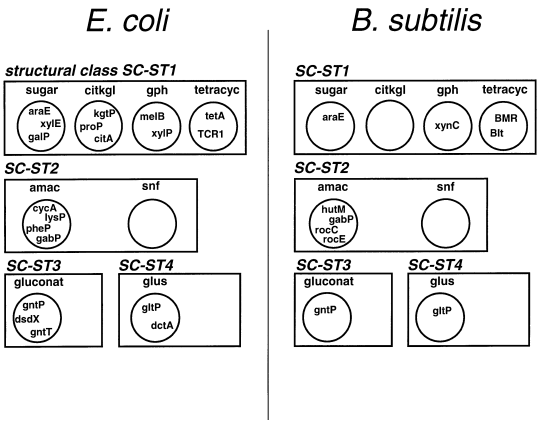


Fig. 5. Structural classification of homologous families of secondary transporters. The classification is based on the analysis of the family hydropathy profiles of 8 secondary transporter families: Sugar, Gph, CitKgl, Tetracyc, AmAc, Snf, Glus and Gluconat [2]. The largest of the structural classes (indicated by the squares) contains 4 different families (indicated by the circles). One class contains two families and the remaining two classes each consist of a single family. Known transporters of *E. coli* and *B. subtilis* that belong to the indicated families are indicated on the left and right, respectively. Some of the transporters are not coded on the chromosome, for instance, the transposon encoded CitA and TetA of *E. coli*. Bmr and Blt are identical to the BMR1bs and BMR2bs sequences in the original definition of the Tetracyc family [2]. Though most families are widely spread throughout nature, the family of Na^+-dependent neurotransmitter transporters (Snf) contained no bacterial homologues.

3. Computational techniques

3.1. Optimal alignment of hydropathy profiles

The power of comparing hydropathy profiles of membrane proteins is that evolutionary relationships can be detected that are more distant than can be detected by comparing amino acid sequences. To make the comparison between hydropathy profiles of amino acid sequences that are not homologous, profiles can be aligned directly [2]. The procedure to do this is based on similar procedures aiming at the alignment of amino acid sequences [16,17]. A hydropathy profile is a one dimensional array of numbers, each representing the average hydrophobicity of the residues in the window that is sliding over the amino acid sequence or, in case of a family profile, over the multiple sequence alignment. To find the optimal alignment of profile a consisting of N windows, $a_1,a_2,a_3,a_4,.....a_{n-1},a_n$, and profile b consisting of M windows, $b_1,b_2,b_3,b_4,.....b_{m-1},b_m$, profile a is converted into profile b. In the conversion, three types of operations are allowed: replacements, insertions and deletions. Each of these operations is associated with a cost. The cost for the replacement of hydrophobicity a_i by b_j equals the absolute difference in hydrophobicity $c_{i,j} = |a_i - b_j|$. The cost for the insertions and deletions that result in the gaps in the final alignment discriminates between creating new gaps and extending existing gaps, $c_{gap} = g + k*h$ in which g is the open gap cost, h the cost for extending a gap and k the number of gaps. The costs for each of these operations accumulates in a cost of conversion for every possible conversion. The optimal alignment is the conversion with the lowest total cost. Computational techniques to find the conversion with the lowest cost make use of recursive algorithms and dynamic programming techniques [18,19]. Though developed for the alignment of amino acid and nucleotide sequences these techniques are equally applicable for the alignment of hydropathy profiles.

Table 2
Search for structural paralogues of membrane proteins with unknown function on the *E. coli* and *B. subtilis* genome

Query	Hit	PDS	Function	Family
Escherichia coli				
b0427	proP	0.138	Proline/betaine transporter	CitKgl
b0486	sdaC	0.124	Serine transporter	
b0845	glpT	0.134	Glycerol-3-phosphate transporter	
b0899	tnaB	0.142	Low affinity tryptophan permease	
b1296	lysP	0.126	Lysine transporter	AmAc
b1690	nanT	0.116	Putative sialic acid transporter	
b1801	caiT	0.122	Probable carnitine transporter	
b2246	cynX	0.134	Cyanate transporter	
b2322	melB	0.129	Melibiose transporter	GPH
b2789	uhpC	0.127	Hexose phosphate uptake regulation	
ydeF	uhpC	0.128	Hexose phosphate uptake regulation	
yhfC	nupG	0.127	Nucleoside transporter	
yhfM	lysP	0.129	Lysine transporter	AmAc
yicM	galP	0.119	Galactose transporter	Sugar
yidT	uhpT	0.124	Hexose phosphate transporter	
yjeM	mtr	0.118	Tryptophan transporter	
yjiJ	narU	0.126	Nitrite extrusion protein	
B. subtilis				
YcsG	lctP	0.119	Putative L-lactate transporter	
YcxA	csbX	0.125	α-Ketoglutarate transporter	
YdeR	araE	0.132	Arabinose transporter	Sugar
YdfA	citM	0.122	Mg^{2+}/citrate transporter	
YdgK	araE	0.141	Arabinose transporter	Sugar
YdhL	araE	0.141	Arabinose transporter	Sugar
YfhI	araE	0.135	Arabinose transporter	Sugar
YfiQ	csbX	0.121	α-Ketoglutarate transporter	
YfkL	csbX	0.123	α-Ketoglutarate transporter	
YhjB	brnQ	0.126	Branched chain amino acid transporter	
YhjI	csbX	0.143	α-Ketoglutarate permease	
YtbD	Bmr	0.110	Multidrug resistance protein	Tetracyc
YuxJ	Blt	0.126	Multidrug resistance protein	Tetracyc
YwbF	araE	0.148	Arabinose transporter	Sugar
YwfF	IolF	0.134	Putative myo-inositol transport	
YxlH	araE	0.124	Arabinose transporter	Sugar
YybF	csbX	0.111	α-Ketoglutarate permease	
YycB	araE	0.144	Arabinose transporter	Sugar

The 'hits' are the proteins of known function of which the hydropathy profile gives the closest match (PDS) to the profile of the proteins of unknown function (query) in the *E. coli* and *B. subtilis* databases. Only hits in the same database are reported. The databases analyzed contained a subset of the membrane proteins on the genomes of *E. coli* and *B. subtilis* [14,15]. For the *E. coli* genome, the annotation used is from version M52. For the *B. subtilis* genome, the annotation of data release R14.2 was used. All gene products annotated as b numbers and y codes were assumed to have an unknown function.

3.2. The PDS and SDS parameters

The profile difference score (PDS) is a measure of the similarity of two hydropathy profiles. The PDS measures the difference in hydrophobicity between the two aligned profiles averaged over all positions. The unit of the PDS is the hydrophobicity unit used to compute the profiles. A low PDS corresponds to similar profiles, but a low PDS is no guarantee for structural similarity. Especially when the number of transmembrane segments in the structure of the two proteins under investigation are different, false positives may occur. Then, the fewer hydrophobic regions (peaks) in one hydropathy profile will distrib-

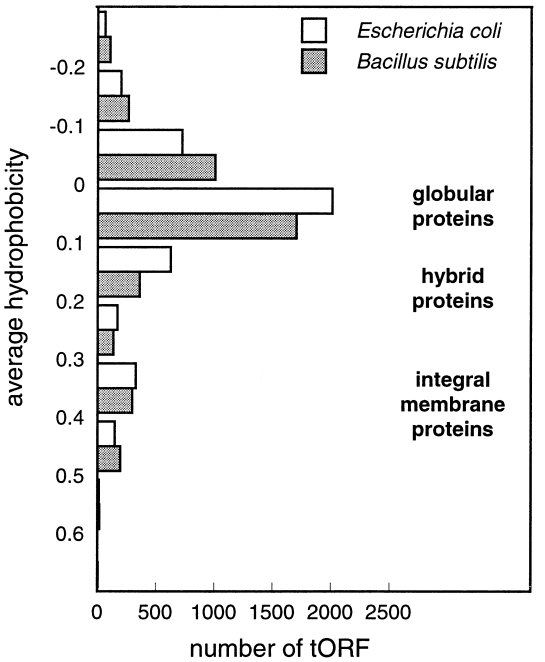


Fig. 6. Average hydrophobicity distribution of the coded sequences on the *E. coli* and *B. subtilis* genomes. The average hydrophobicity of the amino acid side chains of proteins discriminates between soluble proteins and membrane proteins. The hydrophobicity scale used was from Eisenberg [36]. The list of translated ORFs (version M52) of the *E. coli* genome [14] was downloaded from the Web site of the *E. coli* Genome Project at the University of Wisconsin-Madison. The list of translated ORFs (release R14.2) of the *B. subtilis* genome [15] was downloaded from the Subtilist Web site.

ute over the peaks in the other profile to give the 'best fit'. Such alignments are meaningless in terms of structural similarity. A similar situation occurs in amino acid sequence alignments when a short sequence is aligned with a much longer sequence which may result in unrealistically high identity scores. Occasionally, a low PDS is obtained for the optimal profile alignment of proteins with similar numbers of transmembrane segments when a peak in one profile is not present at the same position in the other profile and vice versa. Usually, false positives are easily recognized by visual inspection of the profiles.

The structure divergence score (SDS) measures the divergence of the hydropathy profiles of the members of a family of membrane proteins. The SDS is defined as the difference of the family profile and the individual profiles averaged over all positions and members. The PDS and SDS parameters (defined mathematically in [2]) played an important role in the definition of the structural classes mentioned above (Fig. 5). Structural similarity of two families of membrane protein was inferred from a comparison of the PDS of the optimally aligned family profiles and the SDSs of the two families [2].

4. Genome analysis

4.1. Membrane proteins on the Escherichia coli and Bacillus subtilis genomes

Amino acid sequences coding for integral membrane proteins and globular proteins can be discriminated by the average hydrophobicity of the amino acid side chains. The distribution of the average hydrophobicity of all the translated open reading frames detected on the genome of the Gram-negative bacterium *Escherichia coli* and the Gram-positive bacterium *Bacillus subtilis* shows that most proteins have an average hydrophobicity between 0 and 0.1 (Fig. 6). These are all global proteins. A second maximum in the distribution representing much less proteins is observed between 0.3 and 0.4; these are integral membrane proteins. Hybrid proteins containing both integral membrane bound parts and hydrophilic parts cluster around average hydrophobicities between 0.1 and 0.3. The distribution profiles for the two bacteria are very similar.

The present study focuses on secondary transporters of the 10–14 helix type which have a length of about 450 residues. Therefore, a subset was selected from each genome that contains all amino acid sequences with an average hydrophobicity between 0.2 and 0.5 and with a length between 350 and 550 residues. This resulted in 257 and 215 sequences for *E. coli* and *B. subtilis*, respectively, which amounts to 5–6% of all the sequences on the genomes. In the following sections these two subsets will be used to show how the hydropathy profile alignment technique can be used in the analysis of the data generated in genome sequencing projects.

4.2. Search for structural homologues

A first application of the hydropathy profile

Table 3
Classification in 4 structural classes (SC-ST1–4) of a subset of membrane proteins coded on the *E. coli* and *B. subtilis* genomes

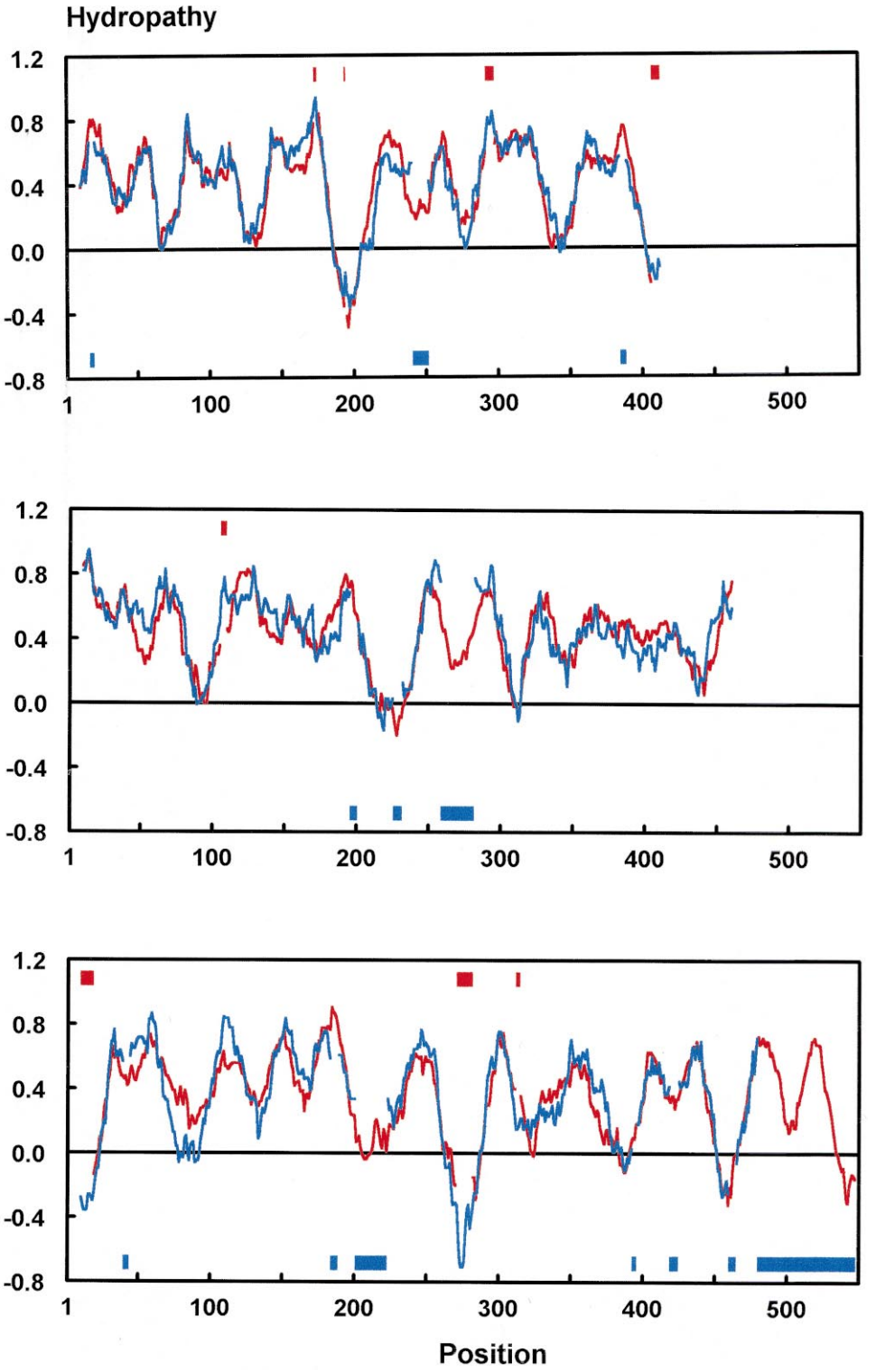
Family[a]	*Escherichia coli*	*Bacillus subtilis*	Function
SC-ST1			
1. SP (Sugar)	araE galP xylE ygcS yaaU ydjE	AraE YdjK YfiG YncC Ywtg YxcC	Arabinose, galactose, xylose symporters
2. DHA (Tetracyc)	bcr emrD mdfA yjiO ydhC yidY yajR yceE ybdA yhfC ydeF	Blt Bmr Mdr Mmr LmrB YusP YvkA YvmA YfiU YhcA YwoD TetB YcnB YceJ YuxJ YitG YwoG YqjV YttB YdgK	Drug antiporters
3. OPA	glpT uhpC uhpT	GlpT	Glycerol-3-P, hexose-P transporter/-receptor
4. OHS	lacY		Lactose symporter
5. MHS (CitKgl)	kgtP proP shiA b1543 yhjE		α-Ketoglutarate, proline, shikimate transporter
6. FGHS	fucP		Fucose symporter
7. NNP	narK narU	NarK	Nitrite exporters
8. NHS	xapB nupG b2098		Nucleoside, xanthosine transporters
9. OFA	yhjX		
10. SHS	nanT yjhB		Sialic acid transporter
11. ACS	ExuT b2789 b4356 yhaU yidT b2246	YcbE YjmG YybO	Hexuronate transporter
12. AAHS	mhpT	YceI	Hydroxyphenylpropionate transporter
13. CP	CynX b1791	YycB	
14. GPH	melB uidB yihO yihP yagG yicJ	YdjD YjmB YnaJ	Melibiose, glucuronide symporters
15.	araJ b1657 ydeA yicM	Ybcl YdhL YtbD YfhI	Transport or processing of arabinose polymers
16.	ydhE	YojI YoeA YpnP YisQ	
17.	yabM yeiO yicK		
18.		CsbX YoaB	α-Ketoglutarate transporter
19.	b1690 b1691	YfkL	
20.	b2322 yhhS		
21.	b2775 yihN		
22.		YddS YdeG	
23.	fsr	YfnC	Fosmidomycin resistance protein
24.	ynfM	YybF	
26.	b1775 ydjE	YyaJ	
27/57 Misc[b]	chaA rfbX yjiJ yhiM ycaD ygeD b0845 b1065 b2046 b2389 b2536	IolF YcxA YbfB YdeR YfiQ YfiS YfkF YfmI YhjI YjcL	Calcium/proton antiporter, hexuronate transporter
		YkuC YlnA YqgE YvqJ YwbF	Putative *o*-antigen transporter
		YwfF YxaM YxiO YxlH YtgP	Putative myo-inositol transporter
SC-ST2			
1. APC (AmAc)	aroP cycA gabP lysP pheP proY b1453 yifK ykfD cadB xasA potE yjdE arcD yhfM yjeH ybaT	AapA GabP HutM RocC RocE YbgF YbxG YdgF YtnA YvbW YbeC YveA YvsH YkbA YecA YfnA YhdG	Amino acid, GABA transporters, cadaverine/lysine, putrescine/ornithine antiporters
2. STP	sdaC tdcC b2845 yhaO yhjV		Serine, threonine transporters
3. ArAAP	mtr tnaB tyrP		Tryptophan, tyrosine transporters
4. SSS	panF putP	OpuE YcgO	Pantothenate, proline transporters

Table 3 (*Continued*)
Classification in 4 structural classes (SC-ST1–4) of a subset of membrane proteins coded on the *E. coli* and *B. subtilis* genomes

Family[a]	*Escherichia coli*	*Bacillus subtilis*	Function
5. LIVSS	brnQ	BrnQ BraB	Branched chain amino acid transporters
6. NCS1	codB allP	YwoE YxlA	Cytosine, allantoin transporters
7. Amt	amtB	NrgA	Putative ammonium transporters
8. NSS (Snf)		YhdH YocR	
9.	b1296 yeeF		
10.		YhjB YodF	
11.	xasA yjeM b0899		Amino acid antiporter
12.	b2392	YdaR	
SC-ST3			
1. GntP (Gluconat)	gntP gntT dsdX b2740 yjhF yjgT	GntP YojA	Gluconate, D-serine transporters
2. Dcu	dcuA dcuB		Dicarboxylate transporters
3/6. Misc[b]	arsB b0621	YhfA YuiF	Arsenical pump subunit B
SC-ST4			
1. DCS (Glus)	gltP dctA b1729	GltP GltT YdbH YhcL	Glutamate, dicarboxylate transporters
2.	ygjU		

[a]Family nomenclature was adopted from [34]. The Sugar, CitKgl, Tetracyc, AmAc, Snf, Gluconat and Glus families referred to in the text are subsets of the SP, MHS, DHA, APC, NSS, GntP and DCS family, respectively.
[b]Listed gene products do not share homology with any of the other gene products on the two genomes analyzed.

alignment technique in genome analysis is the identification of structural homologues of genes that code for membrane proteins of unknown function and for which no sequence homologues with known function are known. Structural identification will indicate the function and possibly the substrate specificity of the unknown protein. The *E. coli* and *B. subtilis* databases of membrane proteins defined above contain 91 and 99 sequences, respectively, that do not have paralogues of known functions within the two databases when a cut-off of 20% sequence identity was used as the detection limit. The hydropathy profiles of these sequences were aligned with all the entries in the databases to search for structural paralogues with a known function. Alignments resulting in a maximum PDS cut-off of 0.15 were selected and the alignments with hydropathy profiles of sequences with known function were visually inspected for false positives. In line with the expectations, hydropathy profile alignment picked up many more paralogues than amino acid sequence alignment. In the *E. coli* data set, 17 of the sequences without paralogues of known function could be shown to be structurally related to sequences with known function. In the *B. subtilis* data set this number was 18 (Table 2). The evolutionary relationship between some of the pairs given in Table 2 was confirmed by a BLAST search [20] of the available databases which identifies sequence similarities in fragments of two sequences. For example, b1801 and caiT of *E. coli* share 26% sequence identity in a stretch of 295 residues ($P = 1.5e{-}77$). Other unknown sequences produced only high-scoring segment pairs with P values over 1.0e−5. For example, yjiJ of *E. coli*, a structural homologue of the nitrite extrusion protein narU, has the highest similarity with an unknown transporter of *Arthrobacter* sp. with a 28% sequence identity in a 35 residues long stretch ($P = 0.00042$). Remarkably, yjiJ has an even lower similarity ($P = 0.023$) to the nitrite extrusion protein NarK of *Helicobacter pylori*. It should be stressed that hydropathy profile alignment identifies the structural class to which a protein belongs. The substrate specificity of the protein with the lowest PDS found in the search can only be an indication of the function of the query protein. To confine the


Hydropathy
1.2
0.8
0.4
0.0
-0.4
-0.8
1
100
200
300
400
500
Position

substrate specificity of the query, it is necessary to classify all the transporters with different substrate specificity into structural classes.

4.3. Structural classification of the membrane proteins

The second application of the hydropathy profile analysis aims at classifying membrane proteins on the genomes in structural classes. The search for structural homologues described above identified many proteins that are members of the Sugar, CitKgl, GPH and Tetracyc families that all belong to the same structural class, SC-ST1 (Table 2, last column). This suggested that proteins belonging to SC-ST1 are abundantly present on the *E. coli* and *B. subtilis* genomes. SC-ST1 is one out of four structural classes of secondary transporters that were identified comparing the family hydropathy profile of 8 different families (Fig. 5) [2]. We have used these family profiles as templates to screen the *E. coli* and *B. subtilis* subsets of membrane proteins for members of the structural classes by the following step wise procedure.

Step 1. The family profiles were aligned with the hydropathy profiles of all the proteins in the databases. Alignments that resulted in a PDS smaller than 0.18 were selected.

Step 2. All selected alignments were inspected visually for false positives. These were removed from the lists.

Step 3. The lists of the Sugar, GPH, Tetracyc and CitKgl that are in the same structural class were pooled in one list. The same was done for the AmAc and Snf lists that are in structural class SC-ST2. The lists of the 4 classes thus obtained were further optimized by the following steps which affected no more than about 25% of the entries.

Step 4. A few entries present in more than one structural class were re-evaluated and assigned to either one of the classes or rejected all together.

Step 5. The proteins in the lists were screened for paralogues in the original databases using a cut-off of 25% overall sequence identity. In case a homologue was found that was not present in the same class-list, the hydropathy profiles of query and hit were re-evaluated resulting in removing the query from the list or adding the hit to the list. The rationale behind this step is that homologous proteins should be in the same structural class.

The results for the *E. coli* and *B. subtilis* subset of membrane proteins are presented in Table 3. The *E. coli* genome and the *B. subtilis* genome code for no less than 116 and 101 proteins, respectively, that cluster in only two structural classes. SC-ST1 contains roughly 150 membrane proteins on the two genomes that are distributed over 57 homologous families. Thirty-one of these families consist of a single gene product that does not have a paralogue or orthologue on the two genomes considered here. The SC-ST1 class is defined by the family profiles of the Sugar, CitKgl, GPH and Tetracyc families and members of these families are abundantly coded on the genomes. The Tetracyc family is the largest family, especially in *B. subtilis* where it contains 20 members. If the unknown sequences represent drug resistance proteins as well, this would emphasize the importance of defense mechanisms for the cell. The original Tetracyc family contained members on the *B. subtilis* genome (Blt, Bmr), but not on the *E. coli* chromosome. The *E. coli* transporters in the Tetra-

←

Fig. 7. Hydropathy profile alignments of secondary transporter family profiles and individual profiles of proteins on the *E. coli* genome. Top: Tetracyc family (red) and the ydhC protein (blue). Though the overall sequence identity between the *E. coli* and *B. subtilis* members of the Tetracyc family is low, the hydropathy profiles of the members from the two bacteria are very similar. The PDS of the alignment was 0.112. The highest sequence similarity of ydhC of *E. coli* with one of the *B. subtilis* proteins is 22% with the Mdr sequence. Middle: Gluconat family (red) and the arsB protein (blue). ArsB is a subunit of an ATP driven arsenate extrusion protein complex that functions as a secondary transporter in the absence of the other subunits. The alignment suggests that the hydrophobic region in arsB between position 240 and 320 in the alignment contains two transmembrane segments which is in agreement with the biochemical evidence [39]. The PDS of the alignment was 0.115. Bottom: The AmAc family (red) and the tdcC protein (blue). Members of the serine/threonine family in SC-ST2 seem to miss the most C-terminal transmembrane segment. The PDS of the alignment was 0.125. All profiles were computed with a window of 19 residues and the optimal alignments were computed using gap costs of 50 for opening a gap (*g*-parameter) and 30 for extending a gap (*h*-parameter).

cyc family have only marginal sequence identity to the *B. subtilis* transporters and their classification in the Tetracyc family is additionally based on the very similar hydropathy profiles (Fig. 7, top). The Sugar and Gph families are represented on both genomes with a total of 12 and 9 members, respectively. The CitKgl family has 5 members on the *E. coli* genome and none on the *B. subtilis* genome. A number of new families in this structural class contain mostly proteins of unknown function, i.e. families 15, 16 and 17. Present on both genomes are members of the family of glycerol-P and hexose-P transporters and receptors (OPA; family 3) and of the family of nitrite extrusion proteins (NNP; family 7). A family of nucleoside and xanthosine transporters (NHS; family 8) is only represented in *E. coli*. Among the proteins that do not have homologues in these two organisms is the well studied lactose permease of *E. coli*, LacY (OHS; family 4) [21].

The second structural class is smaller but still contains a considerable number of members on both genomes. SC-ST2 is represented by 37 members on the *E. coli* genome and 29 members on the *B. subtilis* genome. Twelve families can be discriminated, all of which are represented on both genomes considered here. SC-ST2 is defined by the family profiles of the AmAc and Snf families. The latter contains only transporters from eukaryotic origin, but two homologues are found on the *B. subtilis* genome (NSS; family 8). In contrast, the AmAc family is the largest family with 34 members, distributed equally over the two organisms. The analysis presented here reveals a number of new families that belong to SC-ST2, most notably, transporters for serine and threonine (STP; family 2), aromatic amino acids (ArAAP; family 3), Na^+-dependent symporters (SSS; family 4), branched chain amino acids (LIVSS; family 5) and nucleobase symporters (NSC1; family 6). The amtB and NrgA (Amt; family 7), which are putative ammonium transporters are also in the ST2 structural class.

The two remaining structural classes are considerably smaller than the SC-ST1 and SC-ST2 classes. SC-ST3 mainly contains the members of the Gluconat family (GntP; family 1) with 6 and 2 members on the *E. coli* and *B. subtilis* genomes. New families in this class are formed by two dicarboxylate transporters on the *E. coli* genome (Dcu; family 2) and a family that contains arsB, an arsenical resistance protein with a dual energy coupling mode (Fig. 7, middle). Transport is either driven by ATP when arsB is a component of a multi-subunit complex or by the proton motive force in the absence of the other subunits [22]. Finally two proteins of *B. subtilis* with unknown function belong to this class. SC-ST4 is clearly the smallest structural class in these two organisms. Besides the members of the Glus family (DCS; family 1) that contains bacterial and eukaryotic glutamate transporters, only one other protein on the *E. coli* genome could be assigned to this structural class.

5. Discussion

5.1. Abundance of transport proteins

The hydropathy profile analysis reported here shows that the *E. coli* and *B. subtilis* genomes each code for over 100 sequences that fall in only two structural classes of membrane proteins that, with a few exceptions, are secondary transporters. Two smaller structural classes contain an additional 10 secondary transporters. The classification procedure was rather stringent and it is likely that additional members of especially the SC-ST1 and SC-ST2 classes are in the remaining part of the subset of membrane proteins analyzed. Moreover, in the remaining part (127 and 109 sequences for *E. coli* and *B. subtilis*, respectively) are other known secondary transporters that belong to different structural classes. Clearly, secondary transporters are the most abundant functional type of proteins coded on the genomes of the two bacteria. The large number of secondary transporters and ABC transporters, that form the largest paralogous gene families on the two genomes [14,15,23], emphasizes the importance of communication of the cell with the external world for survival.

5.2. Evolution of secondary transporters

Hydropathy profile analysis discriminates between structural features of membrane proteins. The members of the SC-ST1, SC-ST2, SC-ST3 and SC-ST4 classes are likely to have a different tertiary organ-

ization of the transmembrane α-helices. During evolution, secondary transporters may have evolved from a common primordial gene. If the structurally distinct members of the different classes would have the same ancestor, it is likely that they originate from early gene duplications after which the resulting proteins evolved to different structures. Alternatively, the different classes have emerged from convergent evolution. Possibly, further analysis of both the amino acid sequence and hydropathy profiles of the membrane proteins on the *E. coli* and *B. subtilis* genomes and, most particularly, the comparison with the transporters on the genomes of other organisms allow to discriminate between these two possibilities.

Each of the structural classes of the two bacterial species examined contain paralogues that have diverged so far during evolution that their relation is only evident from hydropathy profiles, i.e. from structural features. In contrast, between the two bacteria there are many orthologues that are much closer suggesting a tight selection pressure or a more recent common ancestor. Examples are the glycerol-3-P transporters GlpT in SC-ST1 that share 60% sequence identity, the arabinose transporters AraE (SC-ST1, 30%), the GABA transporters GabP (SC-ST2, 47%), the branched chain amino acid transporters BrnQ (SC-ST2, 37%), the gluconate transporters GntP/yjhF (SC-ST3, 59%) and the glutamate transporters GltP (SC-ST4, 44%). Though it is possible that transporters for specific substrates are under a more tight selection pressure than transporters for other substrates resulting in independent evolution of these orthologues in the two species, the observation may also be indicative of intensive mixing of the genetic information between different organisms during evolution.

5.3. Two main structural classes

The SC-ST1 and SC-ST2 classes of membrane proteins are widely spread in nature. Originally, the SC-ST2 class contained a family of amino acid transporters (AmAc) and a family of Na^+-coupled neurotransmitter transporters (Snf) [2]. The present analysis shows that already on the genomes of two bacteria many more families can be assigned to SC-ST2. The transporters of SC-ST2 are antiporters or symporters, many of which are Na^+-coupled, with a rather well defined substrate specificity. The substrates are mostly amino acids or amines. In contrast, SC-ST1 is much more diverse and even contains members that are not transporters. UhpC is a receptor specific for glucose-6-P that is part of a two-component signal transduction pathway [24]. SC-ST1 contains uniporters, symporter and antiporters with very different substrate specificity. The largest groups of substrates of the members in SC-ST1 are carbohydrates that are taken up by the cell (symport) and toxic compounds that are extruded from the cell (antiport). Unfortunately, substrate specificity does not always define the structural class to which a transporter belongs. For example, proline transporters are found in both the SC-ST1 and SC-ST2 and citrate transporters are found in SC-ST1 and in two additional structural classes both being represented on the *B. subtilis* genome, the 2-hydroxycarboxylate transporters [25,26] and the Mg^{2+}-dependent citrate transporters [27].

Transporters belonging to either SC-ST1 or SC-ST2 do not necessarily contain the same number of transmembrane segments even though the majority most likely folds as a 12 helix bundle. Well-known is the example of the drug resistance proteins in SC-ST1 that are believed to contain 12 or 14 transmembrane segments (reviewed in [28]). The sialic acid transporter of *E. coli* contains an additional domain that could fold as two extra transmembrane segments in addition to the 12 segments commonly observed [29]. An example of a transporter in SC-ST1 with less than 12 segments is the rhamnose transporter RhaT of *E. coli* that is believed to be a 10 helix bundle [30]. A similar situation is observed in SC-ST2. Again, most proteins are believed to be 12 helix bundles, but the members of the serine and threonine transporters family (STP; family 3) seem to miss the most C-terminally located transmembrane segment (Fig. 7, bottom). Apparently, not all transmembrane segments are equally important for the transport function of the proteins.

5.4. Major Facilitator Superfamily

In the past, a number of homologous families of secondary transporters have been grouped in one superfamily based on sequence motifs present in

members of the families. The Major Facilitator Superfamily (MFS) was originally defined as a group of sugar, drugs and Krebs cycle intermediates transporter families [12,13,31], but has since evolved into a superfamily containing 17 families [32]. The Sugar, CitKgl, Tetracyc and GPH transporter families used here to define SC-ST1 were all part of the families that defined the original MFS. It turns out that SC-ST1 defined by hydropathy profile alignment largely corresponds to the MFS defined in [32], thereby giving this superfamily a similar structure as the common denominator. An important difference between SC-ST1 and the MFS is the ommision of the GPH family [33] from the latter. Though originally suggested to be part of the superfamily because of the presence of characteristic sequence motifs in some of its members, apparently, the GPH family did not fulfil the sequence similarity criteria used in [32]. Hydropathy profile alignment clearly includes the GPH family in the SC-ST1 structural class [2]. The sequence motifs are found in the loops between transmembrane segments 2 and 3 and between 8 and 9. Exactly these loops are very well conserved in the hydropathy profiles of the members of the SC-ST1 families and, in fact, are used as a signature in the visual inspection of the profile alignments. The example of the GPH family shows the high potency of the hydropathy profile alignment technique compared to sequence analysis in showing the relation between distant families of membrane proteins, which is also evidenced by the many more families in SC-ST1 compared to the MFS detected on the genomes of only two bacteria.

5.5. Conclusions and future prospects

Hydropathy profile alignment of membrane proteins is an additional useful technique in the analysis of the large amount of data produced in genome sequencing projects. The technique can detect more distant evolutionary relationships between membrane proteins than can be detected by amino acid sequence analysis and is especially useful when overall amino acid sequence similarities become statistically insignificant. In those cases where sequence similarities are detected only in smaller segments of two sequences [20], the hydropathy profile alignment gives an estimate of the overall similarity of the profiles of two proteins in a graphical representation. Even when the relation between two membrane proteins is not at all evident from the amino acid sequences, the relation may still be evident from the alignment of the hydropathy profiles of the two sequences.

About half of the subset of membrane proteins selected in these studies could be assigned to one of the classes by screening the databases with the family profile of known transporter families. The family profiles were constructed based on criteria defined before that aimed at producing the best fingerprint of the global structure of a family [2]. As new sequences are released almost daily, the composition of the families should be updated continuously to improve the family profiles for screening the databases. Family profiles of additional membrane protein families have to be defined to come to a complete classification of all the membrane proteins on the genomes.

A major goal of functional genomics is the identification of the function of unknown sequences. Classification of the unknown membrane proteins by hydropathy profile alignment may identify the type of protein, for instance secondary transporter, and, to the least, provide a clue about the substrate specificity of the protein. The latter assignment may be improved by a detailed analysis of the differences between the family hydropathy profiles within one structural class. Characteristic features of family profiles may include the number of hydrophobic domains, specific loops, the length of loops, the position of gaps, etc.

The analysis of the *E. coli* and *B. subtilis* genomes reveals a similar distribution of the membrane proteins over the four structural classes. On the other hand, preliminary analysis of the remaining part of the subset of membrane proteins shows that not all structural classes are represented on both genomes. For instance, members of the 2-hydroxycarboxylate transporter family [25] are found on the *B. subtilis* genome, but not on the *E. coli* genome. The complete classification of all membrane proteins of bacteria, archaea and eukaryotes will be helpful in understanding the evolutionary relationship between organisms.

References

[1] Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

[2] Lolkema, J.S. and Slotboom, D.-J. (1998) Estimation of structural similarity of membrane proteins by hydropathy profile alignment. Mol. Membr. Biol. 15, 33–42.

[3] Deisenhofer, J., Epp, O., Miki, K., Huber, R. and Michel, H. (1985) Structure of the protein subunits in the photosynthetic reaction center of *Rhodopseudomonas viridis* at 3 Å resolution. Nature 318, 618–624.

[4] Allen, J.P., Feher, G., Yeates, T.O., Komiya, H. and Rees, D.C. (1987) Structure of the reaction center from *Rhodobacter sphaeroides* R-26: the protein subunits. Proc. Natl. Acad. Sci. USA 84, 6162–6166.

[5] Roth, M., Arnoux, B., Ducruix, A. and Reiss-Husson, F. (1991) Structure of the detergent phase and protein-detergent interactions in crystals of the wild-type strain (strain Y) *Rhodobacter sphaeroides* photochemical reaction center. Biochemistry 30, 9403–9413.

[6] Iwata, S., Ostermeier, C., Ludwig, B. and Michel, H. (1995) Structure at 2.8 Å resolution of cytochrome *c* oxidase of *Paracoccus denitrificans*. Nature 376, 660–669.

[7] Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 Å. Science 272, 1136–114.

[8] Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckman, E. and Downing, K.H. (1990) Model for the structure of bacteriorhodopsin on high resolution electron cryo-microscopy. J. Mol. Biol. 213, 899–929.

[9] Havelka, W.A., Henderson, R. and Oesterhelt, D. (1995) Three-dimensional structure of halorhodopsin at 7 Å resolution. J. Mol. Biol. 247, 726–738.

[10] Von Heijne, G. (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. 5, 3021–3027.

[11] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

[12] Griffith, J.K., Baker, M.E., Rouch, D.A., Page, M.G.P., Skurray, R.A., Paulsen, I.T., Chater, K.F., Baldwin, S.A. and Henderson, P.J.F. (1992) Membrane transport proteins: implications of sequence comparisons. Curr. Opin. Cell Biol. 4, 684–695.

[13] Henderson, P.J.F. (1991) Sugar transport proteins. Curr. Opin. Struct. Biol. 1, 590–601.

[14] Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1462.

[15] Kunst F., Ogasarawa, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, V., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.-K., Codani, J.-J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., OíReilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H. and Danchin, A. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

[16] Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

[17] Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. CABIOS 4, 11–17.

[18] Gotoh, O. (1982) An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.

[19] Hirschberg, D.S. (1975) A linear space algorithm for computing longest common subsequences. Commun. Assoc. Comput. Mach. 18, 341–343.

[20] Altschul, S.F., Gib, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 15, 403–410.

[21] Kaback, H.R. (1996) The lactose permease of *Escherichia coli*: past, present and future. In: Handbook of Biological Physics II: Transport in Eucaryotic and Prokaryotic Organisms (Konings, W.N., Kaback, H.R. and Lolkema, J.S., Eds.) pp. 203–227, Elsevier, Amsterdam.

[22] Kuroda, M., Dey, S., Sanders, O.I. and Rosen, B.P. (1997) Alternate energy coupling of ArsB, the membrane subunit of the Ars anion-translocating ATPase. J. Biol. Chem. 272, 326–331.

[23] Linton, K.J. and Higgins, C.F. (1998) The *Escherichia coli* ATP-binding cassette (ABC) proteins. Mol. Microbiol. 28, 5–13.
[24] Island, M.D. and Kadner, R.J. (1993) Interplay between the membrane associated UhpB and UhpC regulatory proteins. J. Bacteriol. 175, 5028–5034.
[25] Van Geest, M. and Lolkema, J.S. (1996) Membrane topology of the Na^+-dependent citrate transporter of *Klebsiella pneumoniae*. Evidence for a new structural class of secondary transporters. J. Biol. Chem. 271, 25582–25589.
[26] Bandell, M., Ansanay, V., Rachidi, N., Dequin, S. and Lolkema, J.S. (1997) Membrane potential malate (MleP) and citrate (CitP) transporters of lactic acid bacteria are homologous proteins. Substrate specificity of the 2-hydroxy-carboxylate transporter family. J. Biol. Chem. 272, 18140–18146.
[27] Boorsma, A., van der Rest, M.E., Lolkema, J.S. and Konings, W.N. (1996) Secondary transporters for citrate and the Mg^{2+}-citrate complex in *Bacillus subtilis* are homologous proteins. J. Bacteriol. 178, 6216–6222.
[28] Paulsen, I.T., Brown, M.H. and Skurray, R.A. (1996) Proton dependent multidrug efflux pumps. Microbiol. Rev. 60, 575–608.
[29] Martinez, J., Steenbergen, S. and Vimr, E. (1995) Drived structure of the putative sialic acid transporter from *Escherichia coli* predicts a novel sugar permease domain. J. Bacteriol. 177, 6005–6010.
[30] Tate, C.G. and Henderson, P.J.F. (1993) Membrane topology of the L-rhamnose-H^+ transport protein (RhaT) from enterobacteria. J. Biol. Chem. 268, 26850–26857.
[31] Marger, M.D. and Saier, M.H. Jr. (1993) A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. Trends Biochem. Sci. 18, 13–20.
[32] Pao, S.S., Paulsen, I.T. and Saier, M.H. Jr. (1998) Major Facilitator Superfamily. Microbiol. Mol. Biol. Rev. 62, 1–34.
[33] Poolman, B., Knol, J., van der Does, C., Henderson, J.F., Liang, W.-J., Leblanc, G., Pourcher, T. and Mus-Veteau, I. (1996) Cation and sugar selectivity determinants in a novel family of transport proteins. Mol. Microbiol. 19, 911–922.
[34] Paulsen, I.T., Sliwinski, M.K. and Saier, M.H. Jr. (1998) Microbial genome analysis: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J. Mol. Biol. 277, 573–592.
[35] Van Veen, H.W., Venema, K., Bolhuis, H., Oussenko, I., Kok, J., Poolman, B., Driessen, A.J.M. and Konings, W.N. (1996) Drug transport mediated by a novel prokaryotic homologue of the human multidrug resistance P-glycoprotein. Proc. Natl. Acad. Sci. USA 93, 10668–10672.
[36] Eisenberg, D. (1984) Three-dimensional structure of membrane and surface proteins. Annu. Rev. Biochem. 53, 595–623.
[37] Driessen, A.J.M., Fekkes, P. and van der Wolk, J.P.W. (1998) The Sec system. Curr. Opin. Microbiol. 1, 216–222.
[38] Matlack, K.E.S., Mothes, W. and Rapoport, T.A. (1998) Protein translocation: tunnel vision. Cell 92, 381–390.
[39] Wu, J., Tisa, L.S. and Rosen, B.P. (1992) Membrane topology of the ArsB protein, the membrane subunit of an anion-translocating ATPase. J. Biol. Chem. 267, 12570–12576.